1. Title: A efficient method for producing off-line closed captions

Proposed is a process for producing off-line closed captions. Closed captions are the textual transcriptions of the audio track of a TV program, and they are similar to subtitles for a movie show. Captions are generally displayed at the bottom of the screen, and should be synchronized with the audio track. Producing closed captions requires transcribing the spoken words and aligning them along the audio track. Using existing technology, it is a time consuming and costly process.

The method described here allows transcribing the spoken words in a comfortable and easy way for the operator and automates the production of closed captions from the transcribed text. It is used off-line after the audio track has been recorded.

2. Abstract

Our method is a five-step process for producing closed captions for a TV program, subtitles for a movie, or other uses for time-aligned transcripts. An operator transcribes the audio track while listening the recorded material. The system helps him/her to work efficiently and produce precisely aligned captions. The speech rate-control component of the system could be used whenever transcripts are required to be generated from spoken audio.

The first step consists of identifying the portions of the input audio that contain spoken text. Only the spoken parts will be sent to the next module. The other parts can be used to generate non-spoken captions. The second step controls the rate of speech depending on how fast the operator types. This module ensures that the spoken text remains understandable by maintaining a constant pitch [Pat4]. While the operator types, the third module records the time the words were typed in. This provides a rough time alignment for the input text [Pat5]. Then the fourth module re-aligns precisely the text on the audio track. That step reuses the method described in [Pat1]. Finally, the text is broken into captions, similar to a sentence for written text, based on acoustic clues by the fifth module.

3. Inventorship

Jean-Manuel Van Thong
French Citizen
6 Cheviot Road, Arlington, MA 02474, USA
Badge: 197446
Phone: (617) 551-7625
Fax: (617) 551-7650
Cost Center: YAQ
Email: jmvt@crl.dec.com
Manager: R. S. Nikhil

Michael Swain
Canadian Citizen
11 Arden Road, Newton, MA, USA
Badge: 331890
Phone: (617) 551-7627
Fax: (617) 551-7650
Cost Center: YAQ
Email: swain@crl.dec.com
Manager: R. S. Nikhil

Beth Logan
Australian Citizen
353 Harvard Street #22, Cambridge MA 02138, USA
Badge: 331911
Phone: (617) 551-7657
Fax: (617) 551-7650
Cost Center: YAQ
Email: btl@crl.dec.com
Manager: R. S. Nikhil

3.1. WHEN

3.1.1. DATE OF CONCEPTION

The first idea of using audio processing and speech recognition techniques for helping the production of closed caption was discussed by the inventors during fall 1998. The idea of controlling speech rate and segmenting text using acoustic clues was discussed in early April 1999.

3.1.2. DATE OF REDUCTION TO PRACTICE

No experimental tests have been performed yet as of June 1999. We have a complete text/audio aligner system that aligns transcription to the audio track. We plan to implement the closed caption re-aligner module to embed it first into a voice based information retrieval system.

4. The Invention

4.1. PURPOSE

The problem is that of producing efficiently closed captions, or, more-generally, time-aligned transcripts. Closed captions are the textual transcriptions of the audio track of a TV program, and they are similar to subtitles for a movie show. A closed caption (or CC) is typically a triplet of (sentence, time value, and duration). The time value is used to decide when to display the closed caption on the screen, and the duration, when to remove it. Closed captions are either produced *off-line* or *on-line*. Off-line closed captions are edited and aligned precisely along time by an operator in order to appear on the screen at the precise moment the words are spoken. On-line closed captions are generated live, during television newscasts for instance.

Caption can be displayed on the screen in different styles: pop on, roll-up or paint-on. *Pop-on closed captions* appear and disappear at once. Because they require precise timing, they are created during the post-production phase. *Roll-up closed captions* scroll up within a window of 3 or 4 lines. This style is typically used for live broadcasts, like news. In that case, an operator who uses a stenotype keyboard enters the captions live. The *paint-on captions* have a similar style to pop on captions, except they are painted on top of the existing captions, one character at a time.

Captioning a video program is a costly and time-consuming process which costs approximately $1,000 per hour. That includes the whole service from transcription, time alignments, and text editing to make it comfortable to read. Our method could radically cut the cost of producing closed captions.

The number of closed-captioned programs increased dramatically in the US because of new federal laws.

- The landmark Americans with Disabilities Act (or ADA) of 1992 makes broadcasts accessible to deaf and hard-of-hearing.

- The FCC Order #97-279 requires that 95% of all new broadcast programs be closed captioned by 2006.

- The TV decoder circuitry act which imposes all TV 13 inches or larger for sale in the US to have a CC decoder built-in.

In several other countries, legislation requires TV programs to be captioned. On the other hand, DVD have multi-lingual versions and often require subtitles in more than one language for the same movie. Because of the recent changes in legislation and new support for video, the demand for captioning and subtitling has increased tremendously.

The system could also be used whenever time-aligned transcripts are required. Such transcripts are useful for multimedia indexing and retrieval. They can be used to permit the user to precisely locate the parts of the video that are of interest. Emerging application domains include customized news-on-demand, distance learning, and indexing legal depositions for assisting case preparation. Users may access such indexed video and audio via the Internet, using streaming technologies such as RealVideo and RealAudio from RealNetworks.

Aligned transcripts may be used to index video material stored on a digital VCR, either delivered as closed captions integrated into the signal, or delivered separately via an Internet connection. The video appliance project at CRL is building a prototype of such a system.

Recently, a synchronized multimedia standard for the Internet has been developed by the W3C, called SMIL. Among other things, it allows streaming Internet video to be closed-captioned. The SMIL standard is supported by the RealNetworks G2 format. Users can choose whether or not to view the captions or by selecting an option on the ReaPlayer. So in the future, creating multimedia for the World Wide Web may typically involve creating captions, as is currently done for TV.

Finally, the rate-control portion of the system described here would be of value whenever transcripts are required, whether or not the alignment information is needed.

The current systems used to produce closed captions are fairly primitive. They mostly focus on formatting the text into captions, synchronize them with the video, and encode the final videotape. The text has to be transcribed first, or at best imported from an existing file. This can be done in several ways: the typist can use a PC computer with a standard keyboard or stenotype keyboard such as those used by court reporters. At this point of the process, the timing information has been lost and must be rebuilt. Then the closed captions are made from the transcription by splitting the text manually in a word processor. This segmentation can be based on the punctuation, or is made up by the operator. At that point, breaks do make any assumption on how the text has been spoken unless the operator listens the tape while proceeding. The closed captions are then positioned on the screen and their style is defined (italics, colors, uppercase, etc.). They may appear at different locations depending on what is already on the screen. Then the captions are synchronized with the audio. The operator plays the video and hits a key as soon as the first word of the caption has been spoken. At last, the captions are encoded on the videotape, using a caption encoder

In summary, the current industry systems work as follows:

- Import transcription from word processor or use built-in word processor to input text,
- Break lines manually to limit closed captions
- Position captions on screen and define their style,
- Time mark the closed caption manually while the audio track is playing
- Generate the final captioned videotape.

4.2. BACKGROUND MATTER

We first discussed this invention while working on multimedia indexing. One way of indexing multimedia documents is to use the time-aligned transcription of the spoken content. The time stamps associated with each word allow going from the indexed text to the corresponding piece of audio. Closed captions, if they exist, provide you that information. Recently AltaVista partnered with Virage and Compaq Cambridge Research Lab to build a multi-media indexing system. While talking with AltaVista, it was clear that we needed to find solutions to produce time-aligned transcriptions efficiently when they do not exist.

The first solution that comes in mind would be to use speech recognition to generate a time stamped transcription. We are currently running experiments to see how good it would be for indexing. In many cases you would still need an exact transcription and even the best speech recognizers are not accurate enough for that task, especially when the acoustic conditions vary widely. AltaVista looked at having the closed captions produced by a third party: The cost is approximately $1,000 per hour. Considering that this industry is still in its infancy and the methods used are fairly primitive, we started to look for semi-automatic solutions for producing closed captions or more generally time-aligned transcriptions from an audio track.

In some cases however, the transcription already exists, but it is not time stamped. The method described in [Pat2] can be used to accurately align text with the audio track. It uses a speech recognition engine that takes only the words of the input transcription as vocabulary. The output of that module is an aligned transcription where every word is time stamped. In the context of closed caption production, the result can be passed to a caption segmenter module which role is to break the transcription into closed caption segments. The segmenter module is described in Section 4.3.5. Transcribing first, then aligning the text with the audio is a possible approach to produce closed captions. However, valuable time information has been lost while the operator is typing the text in. That information can be used to reduce the complexity of the alignment process and increase its reliability.

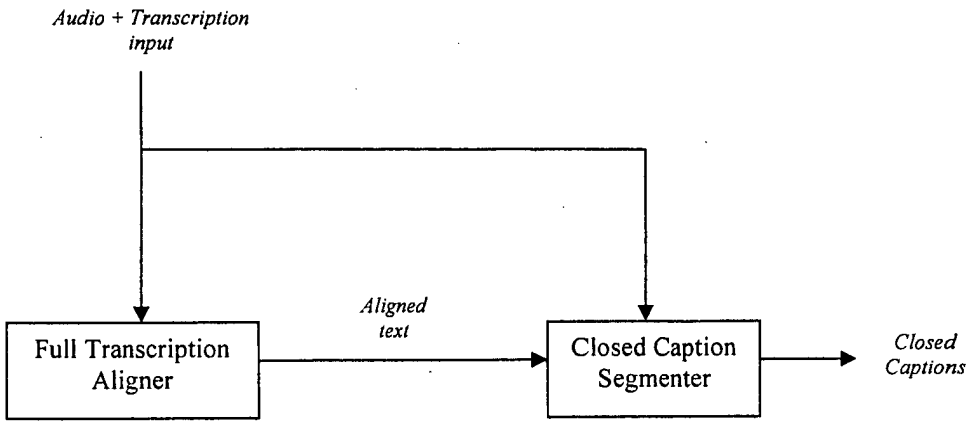

Figure 1: Closed caption production with a full transcription aligner.

If the transcription doesn't exist, we propose here a semi-automatic system that will help the transcriber in his task. The system helps the operator to work efficiently and automates some of the tasks, like segmentation of the captions, and their precise alignment along time. Figure 2 shows the different modules of such a system.

Our method is efficient for the following reasons:

1. The system is efficient and comfortable to use; the operator doesn't have to pause and rewind the recording if it's playing too fast because the system self controls the rate of the speech.

2. The operator can focus on part of the video track that really has to be transcribed (spoken words).

3. We completely eliminate the manual alignment of closed captions along time by capturing time information while typing in the transcription and re-aligning captions as a post-process.

4. Caption segmentation is performed at relevant points in time.

4.3. THE HOWs

Our method is a five-step process for producing closed caption. The system is semi-automatic and requires an operator who transcribes the audio being played. The system helps the operator to work efficiently and automates some of the tasks, like segmentation of the captions, and their precise alignment along time.

The first module, the *audio classifier*, sorts the input audio into different categories: spoken text, music, etc. We are interested in the spoken parts because they need to be transcribed, but also

possibly in particular noise or sound that may need to be captioned. Only the spoken parts are sent to the next module.

The next module, the *speech rate-control module*, controls the rate of speech depending on how fast the text is spoken and/or how fast the operator types. This module ensures that the spoken text remains understandable by maintaining a constant pitch. The audio produced should be time-stamped since a time dependant transformation has been applied to the audio samples. Time stamps will allow the next module to use the proper time scale. This module uses speech recognition techniques at the phoneme level.

The third module, the *time event tracker*, records the time the words were typed in by the operator. This provides a rough time alignment that will be precisely re-aligned by the next module. The recorded time events are mapped back to the original time scale.

The fourth module re-aligns precisely the text on the audio track using speech recognition techniques at the word level.

Finally, the *closed caption segmenter* breaks the text into captions, similar to a sentence for written text, based on acoustic and other clues.

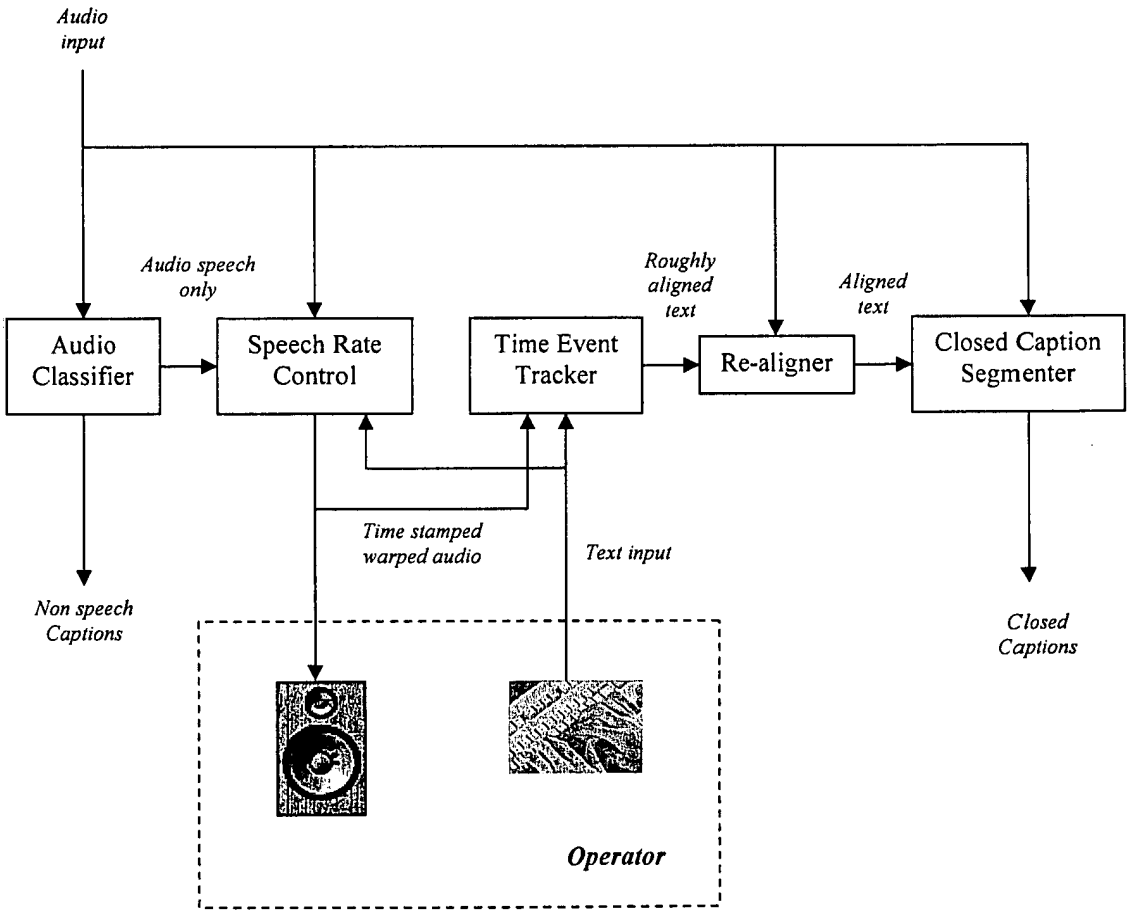

Fig 2: Overall process.

4.3.1. Audio classifier

Before being played to the operator, the audio input may be segmented in order to detect the parts that need to be transcribed, i.e. parts that contains spoken words. This module will allow the operator to concentrate on these parts only and eventually detect sounds or music inserts that also need to be closed captioned (like a barking dog or a train passing by).

This approach is known in the literature as *audio classification* [Erling96]. Numerous techniques can be used to solve the problem. For instance, a HMM or neural net system can be trained to recognize broad class of audio like silence, music, particular sounds, or spoken words. The output is a sequence of *segments*; a segment is a piece of the audio track labeled with the class it belongs to. An example of a speech segmentation system is given in [Hain98].

Note that this module can eventually be integrated with the speech rate control module since that module already performs phoneme recognition. In that case, additional sound or general filler models can be added to the phoneme models in order to capture non-speech audio.

4.3.2. The speech rate control module

This module controls the speech playback rate based on a count of speech units while playing back a recording (speech units could typically be phonemes). It allows adjusting automatically the rate of spoken words to a comfortable rate for the listener. A speech recognizer analyzes a recorded speech stream and produces a count of speech units for a given unit of time. This data, averaged or windowed over a larger unit of time to smooth the results, gives an estimate of the speech rate. A speech-playback-rate adjustment unit uses the computed speech rate to control the speech-playback rate to match a desired rate. The desired speech rate can be a predefined value or depend on an external synchronization, here the keyboard input.

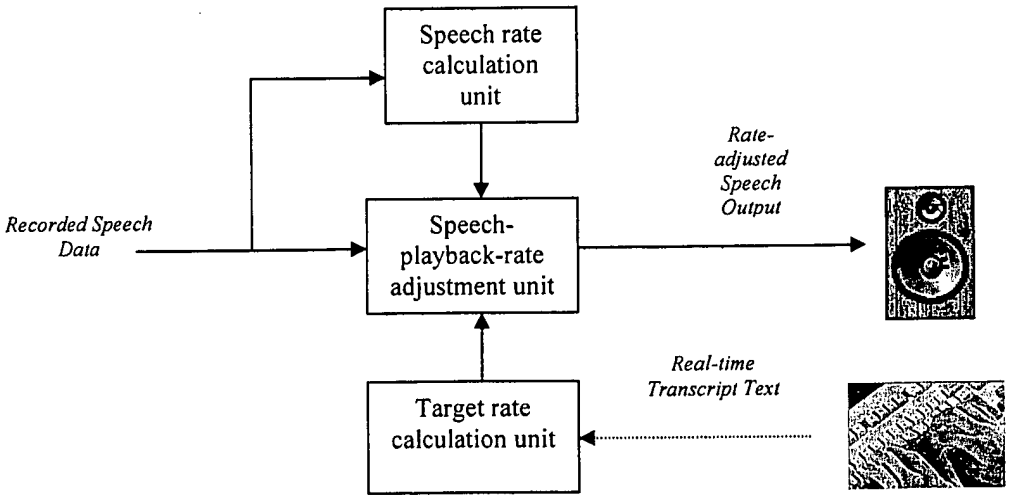

Fig 3: The speech rate control module as described in [Pat4].

This method is completely described in [Pat4].

4.3.3. The time event tracker module

This module automatically links user input with a playback or recording of streaming data via time-stamped trigger events. There are many practical applications of such a system. For example, it can link meeting minutes with an audio recording of the meeting. Here, our closed captioning system uses this module to get a rough alignment between the transcript and the audio or video recording.

This system automatically links user input with a pre-recorded playback or live recording of streaming data via time-stamped trigger events. The system automatically detects pre-defined trigger events, time stamps these events and records time-stamped indices to the events in a master file, in chronological order. User input is thus linked to the streaming data by the nearest-in-time trigger event recorded for the streaming data.

This method is completely described in [Pat5].

4.3.4. The closed captions re-aligner module

This module re-aligns words from the caption stream in order to improve quality of the time marks generated by the time event tracker. Since captions appear on the screen as a group of words determined by the segmenter module, we are only interested in aligning precisely the first and last word of each caption. The time on the first word determines when the caption should appear, and the time mark of the last word determines when the caption should disappear.

The re-aligner uses a combination of speech recognition and dynamic programming techniques where the starting case is audio and closed captions to re-align. The output is a new sequence of captions with improved time alignments. Additional constrains, like video cut time marks or additional delay can be added to improve readability.

This method is completely described in [Pat1].

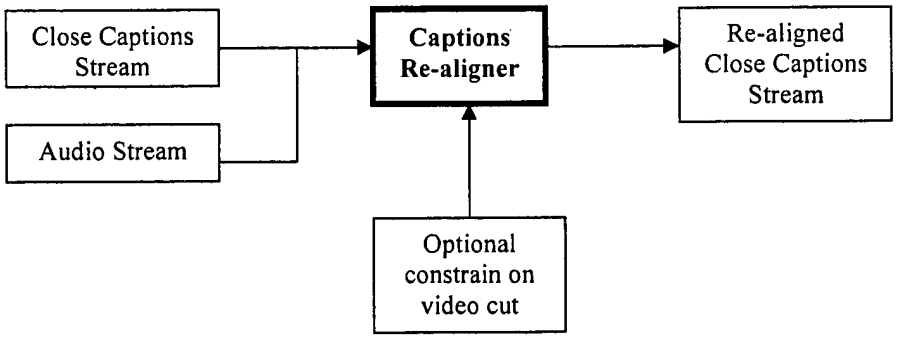

Figure 4: Closed caption re-aligner as described in [Pat1].

4.3.5. The closed caption segmenter module

The closed caption segmenter module takes the stream of roughly aligned text and audio track and finds appropriate break points (silence, breathing, etc.) to segment the text into closed captions. We propose the use of three criteria to find these break points: length of inter-word boundaries; changes in acoustic conditions and natural language constraints. Figure 5 shows the proposed scheme. The individual components are described below.

The output of the closed caption re-aligner module (Section 4.3.4) is time-stamped text. This information is useful to the segmentation process since the length of pauses between words gives an indication of where sentence breaks might be. However, the alignment process is not perfect nor are inter-word pauses necessarily consistent between speakers. Thus we propose also using acoustic and other clues.

Many examples of segmentation schemes based solely on acoustic information exist in the speech recognition literature. For example, [Sie97] describes a segmenter which uses changes in the probability distribution over successive windows of sound combined with energy thresholds to generate segment breaks. The combination of this or a similar scheme with the inter-word pause information lends robustness to the segmentation process.

Additionally, we propose using natural language constraints to verify possible segmentation points. For example, a segment break is unlikely to occur after a "the" or "a". This final piece of information would further increase robustness.

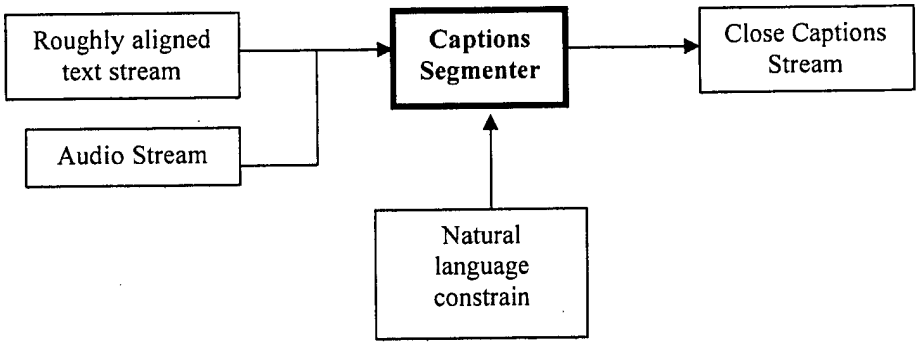

Figure 5: Closed caption segmenter.

4.3.6. Summary

In summary, the process consists of the following steps:

1. classify audio and select spoken parts only, generate non spoken captions if required,
2. operator transcribes the spoken parts of the audio track by using an audio rate control method,
3. time marks are added to the input text using time event keystrokes,
4. re-align precisely the closed captions on the audio track,
5. segment transcribed text into closed captions.

4.4. THE WHYs

There is a multitude of potential applications for a closed captions production system: TV systems, DVD's, multimedia indexing, aids for the deaf, etc. We anticipate these applications will rapidly accelerate the demand for transcribed material (and hence transcription systems) in the future.

In summary, we expect the market for closed captioning to grow substantially. Today, the market for close captioning authoring systems is small because they are currently very expensive, difficult to use, and very labor intensive. Most of the operations are performed manually. At best, the transcription already exists, and an operator has still to align the text onto the audio track. This approach is error prone and depends on how good the person is for that task. Very often, the alignments need to be fine-tuned to meet the quality required. The overall process is costly and time consuming. Cost of closed caption production is roughly $1,000 per hour. With the increasing demand for captioning, the method described could be used efficiently to increase productivity and quality, and reduce production costs.

5. Related Inventions

5.1. PATENTS

[Pat1] *A method for refining time alignments of closed captions*, by JM Van Thong and Pedro Moreno, filed.

[Pat2] *Method for aligning text with audio signals*, by Oren Glickman and Chris Joerg, Application number 08/921,347, filing date 08/29/97, attorney docket number PD25-794. For one of its steps, the present method uses a variant of the algorithm described:

[Pat3] *Automatic indexing and aligning of audio and text using speech recognition*, US patent no 5,649,060. Solve the same problem than [Pat2], but doesn't use recursion for aligning non-recognized words. Also, it doesn't make any assumption on pre-existing time values on word alignments, like closed captions.

[Pat4] *A Speech rate control method using speech recognition*, by JM Van Thong and Davis Pan, to be disclosed.

[Pat5] *A method and system for linking user input with streaming data*, by Davis Pan and Jim Rehg, to be disclosed.

5.2. PAPERS and BOOKS

[Hain98] *Segment Generation and Clustering in the HTK Broadcast News Transcription System,* by T. Hain, S.E. Johnson, A. Tuerk, P.C. Woodland and S. J. Young. In Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[Mor98] *A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments*, by P. Moreno, C. Joerg, JM Van Thong, and O. Glickman. In Proceedings, ICSLP, 1998

[Ribes97] *Automatic Generation of Hyperlinks between Audio and Transcripts*, by J. Robert-Ribes, and R.G. Mukhtar. In Proceedings, EuroSpeech, 1997

[Ribes98] *On the Use of Automatic Speech Recognition for TV captioning,* by Jordi Robert-Ribes. In Proceedings, ICSLP, 1998.

[Rob97] *Inside captioning*, by Gary D. Robson. CyberDawg Publishing, 1997-98.

[Sie97] *Automatic Segmentation, Classification and Clustering of Broadcast News Audio*, by M. A. Siegler, U. Jain, B. Raj and R. M. Stern. In Proceedings DARPA Speech Recognition Workshop, 1997.

[Wold96] *Content-Based Classification, Search, and Retrieval of Audio*, by E. Wold, T. Blum, D. Keislar, and J. Wheaton. IEEE Multimedia v. 3, n. 3, 1996.

5.3. PRODUCTS

The University of Michigan offered a visual communications course where the development of a better close captioning system was major project, see http://www.umich.edu/~viscom/old98/

Several companies commercialize products for transcriptions and close captioning. None of the companies listed below use the method described here.

The CPC Company (Computer Prompting & Captioning Co., 1010 Rockville Pike, Ste 306, Rockville, MD 20852, USA) sells systems for prompting, captioning and subtitling, and provides closed captioning and subtitling service. CPC is the industry leader in closed and open captioning software development. CPC products do not feature the described method, neither any of its components.

Cheetah Systems is a leading supplier of software and systems for court reporting, closed captioning, and real-time litigation. Cheetah sells closed captioning software : TurboCAT, a transcription software, and CAPtivator a CC production software for either off-line or on-line production. CAPtivator seems to capture time codes while typing in the text.

Image Logic (6807 Brennon Lane, Chevy Chase, MD 20815-3255, USA) develop and sells closed captioning software: VidiCaption, DynaCaption, and StudioCaption.

5.4. GENERAL INFORMATION

Many more information on closed captioning can be found at:
http://www.robson.org/capfaq/index.html

5.5. JOINT RESEARCH

This work was solely done at the Cambridge Research Laboratory.

6. Commercialization/Publication

This concept has not been discussed outside of Compaq nor published in any way.

7. Who and Where at Compaq

1. Jean-Manuel Van Thong, jmvt@crl.dec.com, (617) 551-7625
2. Michael Swain, swain@crl.dec.com, (617) 551-7627
3. Beth Logan, btl@crl.dec.com, (617) 551-7657

7.1. YOUR SUPERVISORS

Manager: R. S. Nikhil, CRL Lab Director, (617) 551-7639

7.2. YOUR GROUPS AND BUs

Jean-Manuel Van Thong, Michael Swain and Beth Logan are members of the Compaq Cambridge Research Laboratory (CRL), TCD/CR.

8. Product Use

There are currently no Compaq products embodying this invention. This invention is part of Corporate Research's effort to develop new fundamental technologies for Compaq.

9. Signatures

Jean-Manuel Van Thong

Michael Swain

Beth Logan